STORAGE AND RETRIEVAL OF INFORMATION FROM INSECT SPECIMENS [1]

Ross H. Arnett, Jr.[2]

The crux of taxonomy is the name of the organism, for with it, whether accurately or inaccurately determined, one stores, associates, and retrieves information. The name, therefore, serves as the turnstyle, and often the bottleneck, of information flow. Taxonomists long ago devised their own rigid system for uniformly storing information for easy retrieval, i.e., the binomial system of nomenclature. The storage procedure is controlled by the generally accepted International Code of Zoological Nomenclature (hereafter referred to as the Code). The *only* bridge between the system and biological data is the type specimen, the holotype, the lectotype, or the neotype.

Much confusion in the presentation of taxonomic data is due to the failure to conceive of the Code as an information storage and retrieval program. The program requires only name data to operate, but too frequently the input lacks accurate and detailed information associated with names and specimens, and fails to associate biological information with names correctly associated with types.

This paper deals specifically with the problems resulting from the use of insect specimens stored in collections as a source of information. I hope that it will stimulate improvement of collections and the more critical use of the data obtained from them.

The procedure discussed below concerns three stages termed: input, flow, and output. Input is concerned with the gathering and processing of taxonomic documents. Flow involves the sorting and storage of these documents through the use of names and type specimens. Output from the properly functioning system returns these information documents associated with the correct name. Systematists then use this information.

Today's possibilities for scientific crescendos through information storage and retrieval using data processing machines makes it imperative for one to understand some of the theory and nature of information storage. This introduction is not intended to be a source for storage and retrieval

[1] Accepted for publication June 21, 1969.

[2] Department of Entomology, Purdue University, Lafayette, Indiana 47907.

technique, but rather a discussion of the nature of insect specimens and associated data in terms of modern information theory. Because the systematic entomologist depends upon others for many of his data, the validity of the information he must use should be examined.

The problem

Errors in labelling, distortions through preservation, and lack of detailed data aossciated with the specimens, contribute errors in systematic studies. An awareness of this will serve to improve the recording of data in the future as well as to help the proper handling of existing information. Due mainly to the requirements of the Code, taxonomists are prone to publish too much information too quickly, well before it has been processed sufficiently for use by others. Therefore, a large amount of erroneous information may be passed from publication to publication as fact, for the Code has no article that requires that published information be demonstrable as fact! A statement printed on a locality label and attached to a specimen is not sufficient evidence of data validity.

Two terms are frequently used as if they were synonyms: data and information. Data pertains to facts or statistics, either historical, or derived from calculations or experimentation. Data may be processed in a computer and retrieved in a form that differs from that stored. Information refers only to documents communicated or received concerning a particular fact or circumstance. Information is stored and retrieved without change. We are concerned here only with information processing.

The scientific method requires that observations and experiments be repeatable. This means that, barring catastrophies of nature and man, samples of extant species, and information associated with these samples can be re-gathered by anyone given the proper circumstances. Failure to gather these data does not necessarily invalidate the original data. It may mean, however, that the original information was too scanty to meet scientific requirements. The time now has come for the development of a set of standards for acceptable specimens and associated data. Specimens not meeting these standards should be disregarded except under unusual circumstances.

Input.—Taxonomic input information consists of documents in the form of specimens, label data, observational and experimental data either associated with specimens or with names, and published information compiled from these data, the "raw" data of taxonomy. These are taken in order for further study and consideration to lead to a better system of information storage.

Single specimens in a collection have a 50% chance of having incorrect labels. Even a series of specimens bearing exactly the same label data may be incorrectly labelled. With the collection of two or more specimens in a single locality, one fact is established: some morphological variation of the species. When still more samples have been gathered in the same locality by the same person or different persons at other times or during the same season of other years, the chance of incorrect labelling is insignificant. Once the same species has been collected at other localities, the validity of the theory of the existence of the species is established.

Traditionally, collecting is a random but biased procedure. Seldom has the taxonomic entomologist set out in advance to discover the distribution of a population or a species. This is, like labelling, another weak part of our information gathering system because we are willing to accept the data we have. We grasp these meager data and rush them to the printer to be recorded as a taxonomic "first" without consideration of their true significance! Carefully planned collecting would improve even the first stages of taxonomy.

Insect specimens, to be scientifically useful, should represent samples of a breeding population or deme. This cannot be determined at the time of their capture, but if properly preserved and associated with the required data, this may make possible the determination and circumscription of the population at least as it existed for a particular period of time, recognizing, of course, the ebb and flow of populations. The addition of the collecting date tells others that the particular stage occurs during a particular season. The addition of the collector's name could lead the user to other records by the association of the collector with the specimens. Such records should be directly tied into the data label by reference, so the collector's name becomes more of a bibliographic citation than one of historical interest.

Unlike other scientists, insect taxonomists depend almost entirely upon stored preserved specimens. They rarely work with living cultures either in the field or in the laboratory. There are no insect zoos, no "type cultures," and no surveys of the "vanishing herds." It seems logical to assume that great scientific progress could be made if living material were included as an information source. Further consideration of this is beyond the scope of the present analysis except insofar as it explains the need for specimens.

Certain fields of organismic investigation no longer depend upon stored specimens except under special conditions, because these groups can be recognized to species and even deme without the study of preserved material. Birds of many species can be observed and accurate reports made without the need to collect and preserve specimens. Many other groups

are well enough known so that experiments, as with many vascular plants for example, may be performed without the need to first work with the preserved specimens. In some groups, such as bacteria, stored specimens are of little value and the specialist must depend upon descriptions. One can use a manual or a similar source for the correct identification of the species of well known groups, as in the case of many economically important species. The information obtained from these observations or experiments is then stored for later retrieval by calling for the information by scientific name only. It follows then, that the primary reason for storing information in the form of specimens is to study the morphology of the species so that observational and experimental data may be associated with a species. Once a group is well enough known taxonomically this is no longer necessary.

Flow.—The processing of the information documents, or the flow of information through the system, involves the indexing of these documents. Taxonomists do this by making identifications. They use as the point of reference the holotype specimen, or its substitute. Through this name association, the documents are stored and retrieved.

Once a specimen has been accurately identified, thus providing a name, the most important step in the preparation of information storage has been completed. Accurate means should be devised to positively associate individual specimens with information because identification errors are made and the concept of the species involved may change. If data and specimens cannot be associated individually, the data are useless. This is true of any system. The need for a standardized format for information association has become acute. As the volume of data mounts, a higher percentage becomes inaccessible.

Output.—The system returns information documents only by the correct use of the name as the index. The information retrieved describes the variation, distribution, and biological knowledge about the species. Several uses of retrieved information have been pointed out previously (Arnett, 1967a). These are repeated here with the hope of encouraging a more general awareness of the need for this kind of information.

1. *Holotypes and topotypic populations.* A single specimen, we have concluded previously, does not provide enough information for scientific treatment. Holotype specimens are no exception and it has never been claimed that they served anything other than nomenclatural purposes. This being so, there seems to be no valid scientific reason for continuing to treasure poor holotypes, and there is no practical reason why neotypes should be restricted to replacing lost or destroyed holotypes, while retaining useless originals. As the taxonomist attempts to associate data with

names he may find a hopeless impass because the holotype does not provide the needed data.

Accepting the premise that the scientific method requires repeatability, it follows that specimens of extant species should be recoverable from the field. One way to demonstrate this is to match holotypes with specimens from the field. If, for example, a holotype lacks specific locality information, but can be matched with specimens from a restricted locality, that is, it agrees almost exactly with another variant so that there is little doubt that the extant deme represents a breeding population containing the same variants as the holotype, that population may be regarded as topotypic regardless of whether it is now at the same locality as it was when the holotype specimen was collected. Specimens studied from that population may then supply biological information attributable to the holotype. With these as the anchor for the name, it is a simple matter to continue studies on related populations to help understand the variation and perhaps even the speciation of the group.

Further study of the area surrounding the restricted "topotypic" population, and the gathering of information about these areas may serve to circumscribe a breeding population. It follows that by so doing, other such populations may be circumscribed, making possible the study of isolates if such exist. The holotype then serves as the link between biological data and the information storage and retrieval system.

2. *Variation, distribution, and biological information.* From the input information now associated with the proper species, detailed records of host, habitat, and habits for each sample may add valuable data to show variations within a single population: seasonal, genetic, environmental, and other variations.

Habitat descriptions that include weather data, altitude, slope, vegetation, and other information about the effective environment may be useful not only for the specimens collected at the time of the visit, but for the association of other collections made at other times when these details were not noted.

Information output is reused, added to, and stored back into the system. It is superfluous to say that the proper association of data and species is of paramount importance in the success of any information storage and retrieval system. It is not superfluous to emphasize that the only purpose of the Code is to provide an efficient program for the storage and retrieval of this information. Once a less cumbersome system has been devised, the Code is obsolete. Such a system is at least theoretically possible, so the Code is at least theoretically obsolete. We should work to develop a new system before someone else does it for us (see Mittman, 1968, for a possible system).

New methods possible

Storage and retrieval of information documents can be easily and cheaply computer programmed. Several programs are now available. We could formalize a program for insect label data, but only when some general agreement has been reached with a large body of our taxonomists. Methods have been devised for the rating of the efficiency of an information storage and retrieval system (Lancaster, 1968). We are learning that no existing system is 100% efficient, but the gaining of the last few percentage points to make a system 100% efficient are too expensive in time and money with too little significance in results to make them a worthy goal. The trouble in the past has been in attempting to establish absolute priority for names thereby implying 100% retrieval. The Code has only recently relaxed this requirement by providing for the use of some names that may not be the oldest for the species.

The ability to store information in a system using computer equipment opens the way for new techniques and changes our views on many traditional procedures. For example, one of the primary reasons for the publication of data by printing is to file them in the libraries of the World where the information may be readily and rapidly retrieved. Here these data lie waiting to be used. In such cases it is doubtful if there is even a 1.0% retrieval. However, when the same data are stored in a central bank, if the one "copy" is used once, 100% retrieval is achieved. Time delay disappears once the user is hooked into a chain of data banks. Such a chain will become a reality soon. Thus, it is no longer necessary for many kinds of data to be reproduced by printing. This applies to much or all of the data recorded on insect data labels.

Many branches of biology have been able to quantify working data through the use of various measuring and recording instruments. Little attempt has been made to do the same for data associated with insect specimens. Although insect collections still bear the classical three-line data label—the place, the date, and the collector's name, some improvement has been made by a majority of collectors since the early 1940's. They have been careful to record localities more precisely, and most have included information that might be termed "ecological," but little else. Because of the requirements of the Code, specimens have "historical" value, so most specimens, once information about them has been recorded in print, must be kept regardless of their condition or sparsity of data.

Storage of undigested raw data as printed matter is wasteful. For example, the mere listing of locality records from insect specimen labels, arranged alphabetically or geographically, is still raw data. When printed and stored, ready for use by the next researcher, it is he who must do the

actual research by interpretating as best he can the significance of the data. Lists of temperature recordings taken during some physiological experiment are comparable data and would not be published. We forget that we write primarily for the use of the non-specialist, not the one other specialist on our group. The work should be done for the user, not as a mere appraisal of the current status of the research, but as a finished working tool. The progress reports can be better distributed to the specialists. However, if a list of collecting sites is recorded, especially when it tells others the season during which to collect, what to expect, and provides information about the habitat, then a useful contribution is made.

Original descriptions and first revisions are information compiled from specimens and label data. These data are certainly most suitable information for punch cards or tape storage. The method of storing data is presently governed by the Code, but this need not be restrictive, for the Code is vague in its wording on this matter, and its writers were not considering the changes brought about by computer technology (for a discussion of how information may be treated validly under the Code and at the same time prepared for punch card storage, see Arnett, typescript on "Data Documents").

Endless amounts of information might be gathered and prepared for storage. A simple means of storing this might be to prepare a punch card for each information class, the cards coded to species and type of information. Arranged and stored, these may be changed at any time and are available for immediate printout. The following list suggests several information classes, some of which might be further subdivided:

1. Species card with classification code.
2. Preserved specimen (as individuals) cards keyed to lot records.
3. Literature reference cards (for taxa).
4. Locality data cards.
5. Field record cards.
6. Experimental data cards.
7. Measurement (individual) cards keyed to 2, above.
8. Ecological information cards (this item can be broken into many parts).
9. Photograph record cards keyed to 4 and 5 above.
10. Literature reference cards for ecological data and other information.
11. Cards showing the collection in which each vouchered specimen is stored, also keyed to 2 above.

The format for these cards, and indeed, the entire system, is being studied in several laboratories (Arnett, 1967b) to devise a functional arrangement.

Unfortunately little information on associated data other than that obtainable from the "standard" locality label is readily available with most collections. Some collections, particularly those closely connected with projects of an economic nature, have lot record systems so that many specimens have an abundance of associated information. Without doubt a great amount of data not keyed to the specimens is available from individual collectors. If collectors would prepare these data for distribution with loaned specimens, their research value would be increased greatly.

Meanwhile our current system should be reexamined for deficiencies. Some of the most obvious are summarized here with the hope that future records will be improved.

1. Lack of accurate locality information is a major deficiency. Enough information should be available so that one might return at any time to the exact collecting site to obtain additional specimens.

2. Since much more material is collected than is used by the collector, care should be taken to make sure information is gathered to render the specimens useful to others. For example, host records may be needed, and without them the specimens are merely useless additions to an already bulging collection. I believe these are often described as common "trash." Yet such material might be turned easily into valuable information records by the simple addition of the proper data.

3. The need for enough information so that the observation can be repeated has been mentioned already. The lack of these data is as frequent as the lack of precise locality data.

4. Observers fail to record in notebooks or by photograph many readily available information bits that might easily enrich a storage bank, not only by adding to the file on the particular species under study, but for other species as well. For example, ecological information on a type locality might be valuable data for use in sharpening the concept of the species even if additional specimens were unavailable at the time the observations were made. One should never presume that the information he may record is already available.

The continued collection of light attracted specimens, and similar mass collection procedures will serve as a valuable source of information for some time to come. Most of the specimens borrowed from the collections will be samples of this nature. With the modern techniques of information storage and the easy means this provides for retrieval, it seems clear that more detailed data will be associated with a greater percentage of material, readily available to all, in the very near future.

Literature Cited

Arnett, R. H., Jr. 1967a. Locality and data labels for insects. Ann. Ent. Soc. America, 60: 1111–1112.

——. 1967b. Pilot laboratory: a center for the study of insect populations. Purdue University, Lafayette, 21 pp.

——. 1969. Data documents, in typescript.

Lancaster, F. Wilfrid. 1968. Information retrieval systems. John Wiley and Sons, N. Y., xiv + 222 pp.

Mittman, B. 1968. INFOL for the CDC 6400, Information storage and retrieval system. Northwestern University, Evanston, vii + 10 sections.

The Entomologist's Record

To encourage the publication of concise and useful new distribution records, corrections of previously published erroneous records, misidentifications, short field notes, and current news items about entomologists, amateur and professional, entomology departments and museums, prompt (monthly) publication is offered in this department.

A paper "computer" for Entomologists with limited recall.—From time to time I find it necessary to locate in once-read and since-forgotten articles information dealing with fireflies or other luminescent organisms. For the retrieval of this information I use what essentially amounts to a multiple-drawer card file housed in a single notebook. Since I started using this system my memory has given me much less trouble or concern. In addition to its data retrieval potential, the main advantage of this "paper computer" is that once in operation it requires but a modicum of attention to keep it going. This system has distinct advantages over notched card systems: it does not cost several cents per citation, or several dollars for a puncher; cards aren't turned, stacked, poked, shaken, restacked or bent out of shape; when entering data there are no annoying snaps or pops to acoustically stimulate irritable librarians; and the completed data storage file can be Xeroxed for insurance or by students who wish to take it along when they leave for positions elsewhere.

Figure 1 shows a sample page from the notebook. It contains 29 references and adjacent to each are 20 code letters, each of which represent some aspect of luminescent organisms in which I am interested and wish to isolate for retrieval. The first entry is interpreted as follows: McDermott in 1914 in volume 10 of some journal (a single master card file on 5" × 8" cards gives the complete citation) published an article that deals with (right-hand *margin*) Elateridae (E), Annelida (A), Lampyridae (+) (if no letters appear in this margin the article deals only with Lampyridae), Diptera (D), and Phengodidae (P). The black dots indicate the following: glowing (G) and flashing (F) lampyrids are discussed, larvae and/or lifecycles (L), behavior (B), evolution and/or function (E), and distribution and/or zoogeography (D). Other categories not marked for McDermott's paper include taxonomy (T), Ecology (E in letter group 4), physiology/biochemistry (P in letter group 4), flight and light emission in adults (F in letter group 2), and reproductive isolation (L in letter group 1). Small embellishments of the dots permit more subtle coding distinctions. For example, the dot with the cross over it that appears over the E in letter group 3 (predator/prey)